ANALYSING POLYNUCLEOTIDE SEQUENCES

This is a CIP of Serial No. 07/573,317 filed September 28, 1990.

1. INTRODUCTION

Three methods dominate molecular analysis of nucleic acid sequences: gel electrophoresis of restriction fragments, molecular hybridisation, and the rapid DNA sequencing methods. These three methods have a very wide range of applications in biology, both in basic studies, and in the applied areas of the subject such as medicine and agriculture. Some idea of the scale on which the methods are now used is given by the rate of accumulation of DNA sequences, which is now well over one million base pairs a year. However, powerful as they are, they have their limitations. The restriction fragment and hybridisation methods give a coarse analysis of an extensive region, but are rapid; sequence analysis gives the ultimate resolution, but it is slow, analysing only a short stretch at a time. There is a need for methods which are faster than the present methods, and in particular for methods which cover a large amount of sequence in each analysis.

This invention provides a new approach which produces both a fingerprint and a partial or complete sequence in a single analysis, and may be used directly with complex DNAs and populations of RNA without the need for cloning.

In one aspect the invention provides apparatus for analysing a polynucleotide sequence, comprising a support and attached to a surface therof an array of the whole or a chosen part of a complete set of oligo nucleotides of chosen lengths, the different oligonucleotides occupying separate cells of the array and being capable of taking part in hybridisation reactions. For studying differences between polynucleotide sequences, the invention provides in another aspect apparatus comprising a support and attached to a surface therof an array of the whole or a chosen part of a complete set of oligonucleotides of chosen lengths comprising the polynucleotide sequences, the different

oligonucleotides occupying separate cells of the array and being capable of taking part in hybridisation reactions.

In another aspect, the invention provides a method of analysing a polynucleotide sequence, by the use of a support to the surface of which is attached an array of the whole or a chosen part of a complete set of oligo nucleotides of chosen lengths, the different oligonucleotides occupying separate cells of the array, which method comprises labelling the polynucleotide sequence or fragments thereof to form labelled material, applying the labelled material under hybridisation conditions to the array, and observing the location of the label on the surface associated with particular members of the set of oligonucleotides.

The idea of the invention is thus to provide a structured array of the whole or a chosen part of a complete set of oligonucleotides of one or several chosen lengths. The array, which may be laid out on a supporting film or glass plate, forms the target for a hybridisation reaction. The chosen conditions of hybridisation and the length of the oligonucleotides must at all events be sufficient for the available equipment to be able to discriminate between exactly matched and mismatched oligonucleotides. In the hybridisation reaction, the array is explored by a labelled probe, which may comprise oligomers of the chosen length or longer polynucleotide sequences or fragments, and whose nature depends on the particular application. For example, the probe may comprise labelled sequences amplified from genomic DNA by the polymerase chain reaction, or a mRNA population, or a complete set of oligonucleotides from a complex sequence such as an entire genome. The end result is a set of filled cells corresponding to the oligonucleotides present in the analysed sequence, and a set of "empty" sites corresponding to the sequences

which are absent in the analysed sequence. The pattern produces a fingerprint representing all of the sequence analysed. In addition, it is possible to assemble most or all of the sequence analysed if an oligonucleotide length is chosen such that most or all oligonucleotide sequences occur only once.

The number, the length and the sequences of the oligonucleotides present in the array "lookup table" also depend on the application. The array may include all possible oligonucleotides of the chosen length, as would be required if there was no sequence information on the sequence to be analysed. In this case, the preferred length of oligonuclectide used depends on the length of the sequence to be analysed, and is such that there is likely to be only one copy of any particular oligomer in the sequence to be analysed. Such arrays are large. If there is any information available on the sequence to be analysed, the array may be a selected subset. For the analysis of a sequence which is known, the size of the array is of the same order as length of the sequence, and for many applications, such as the analysis of a gene for mutations, it can be quite small. These factors are discussed in detail in what follows.

2. OLIGONUCLEOTIDES AS SEQUENCE PROBES

Oligonucleotides form base paired duplexes with oligonucleotides which have the complementary base sequence. The stability of the duplex is dependent on the length of the oligonucleotides and on base composition. Effects of base composition on duplex stability can be greatly reduced by the presence of high concentrations of quaternary or tertiary amines. However, there is a strong effect of mismatches in the oligonucleotides duplex on the thermal stability of the hybrid, and it is this which makes the technique of

hybridisation with oligonucleotides such a powerful method for the analysis of mutations, and for the selection of specific sequences for amplification by DNA polymerase chain reaction. The position of the mismatch affects the degree of destabilisation. Mismatches in the centre of the duplex may cause a lowering of the Tm by 10°C compared with 1°C for a terminal mismatch. There is then a range of discriminating power depending on the position of mismatch, which has implications for the method described here. There are ways of improving the discriminating power, for example by carrying out hybridisation close to the Tm of the duplex to reduce the rate of formation of mismatched duplexes, and by increasing the length of oligonucleotide beyond what is required for unique representation. A way of doing this systematically is discussed.

3. ANALYSIS OF A PREDETERMINED SEQUENCE

One of the most powerful uses of oligonucleotide probes has been in the detection of single base changes in human genes. The first example was the detection of the single base change in the betaglobin gene which leads to sickle cell disease. There is a need to extend this approach to genes in which there may be a number of different mutations leading to the same phenotype, for example the DMD gene and the HPRT gene, and to find an efficient way of scanning the human genome for mutations in regions which have been shown by linkage analysis to contain a disease locus for example Huntington's disease and Cystic Fibrosis. Any known sequence can be represented completely as a set of overlapping oligonucleotides. The size of the set is $N - s + 1 \simeq N$, where N is the length of the sequence and s is the length of an oligomer. A gene of 1 kb for example, may be divided into an overlapping set of

around one thousand oligonucleotides of any chosen length. An array constructed with each of these oligonucleotides in a separate cell can be used as a multiple hybridisation probe to examine the homologous sequence in any context, a single-copy gene in the human genome or a messenger RNA among a mixed RNA population, for example. The length s may be chosen such that there is only a small probability that any oligomer in the sequence is represented elsewhere in the sequence to be analysed. This can be estimated from the expression given in the section discussing statistics below. For a less complete analysis it would be possible to reduce the size of the array e.g. by a factor of up to 5 by representing the sequence in a partly or non-overlapping set. The advantage of using a completely overlapping set is that it provides a more precise location of any sequence difference, as the mismatch will scan in s consecutive oligonucleotides.

4. ANALYSIS OF AN UNDETERMINED SEQUENCE

The genomes of all free living organisms are larger than a million base pairs and none has yet been sequenced completely. Restriction site mapping reveals only a small part of the sequence, and can detect only a small portion of mutations when used to compare two genomes. More efficient methods for analysing complex sequences are needed to bring the full power of molecular genetics to bear on the many biological problems for which there is no direct access to the gene or genes involved. In many cases, the full sequence of the nucleic acids need not be determined; the important sequences are those which differ between two nucleic acids. To give three examples: the DNA sequences which are different between a wild type organism and one which carries a mutant can lead the way to isolation of the relevant gene; similarly, the sequence differences between a cancer cell and its

normal counterpart can reveal the cause of transformation; and the RNA sequences which differ between two cell types point to the functions which distinguish them. These problems can be opened to molecular analysis by a method which identifies sequence differences. Using the approach outlined here, such differences can be revealed by hybridising the two nucleic acids, for example the genomic DNA of the two genotypes, or the mRNA populations of two cell types to an array of oligonucleotides which represent all possible sequences. Positions in the array which are occupied by one sequence but not by the other show differences in two sequences. This gives the sequence information needed to synthesise probes which can then be used to isolate clones of the sequence involved.

4.1 ASSEMBLING THE SEQUENCE INFORMATION

Sequences can be reconstructed by examining the result of hybridisation to an array. Any oligonucleotide of length s from within a long sequence, overlaps with two others over a length s-1. Starting from each positive oligonucleotide, the array may be examined for the four oligonucleotides to the left and the four to the right that can overlap with a one base displacement. If only one of these four oligonucleotides is found to be positive to the right, then the overlap and the additional base to the right determine s bases in the unknown sequence. The process is repeated in both directions, seeking unique matches with other positive oligonucleotides in the array. Each unique match adds a base to the reconstructed sequence.

4.2 SOME STATISTICS

Any sequence of length N can be broken down to a set of $\sim$ N overlapping sequences s base pairs in length. (For double stranded nucleic acids, the sequence complexity of a sequence of N base pairs is 2N, because the two strands have different sequences.

but for the present purpose. this factor of two is not significant). For oligonucleotides of length s, there are 4^s different sequence combinations. How big should s be to ensure that most oligonucleotides will be represented only once in the sequence to be analysed, of complexity N? For a random sequence the expected number of s-mers which will be present in more than one copy is

$$\mu_{>1} \approx 4^s(1 - e^{-\lambda}(1 + \lambda))$$

where

$$\lambda = (N - s + 1)/4^s$$

For practical reasons it is also useful to know how many sequences are related to any given s-mer by a single base change. Each position can be substituted by one of three bases, there are therefore 3s sequences related to an individual s-mer by a single base change, and the probability that any s-mer in a sequence of N bases is related to any other s-mer in that sequence allowing one substitution is $3s \times N/4^s$. The relative signals of matched and mismatched sequences will then depend on how good the hybridisation conditions are in distinguishing a perfect match from one which differ by a single base. (If 4^s is an order of magnitude greater than N, there should only be a few, 3s/10, related to any oligonucleotide by one base change.) The indications are that the yield of hybrid from the mis-matched sequence is a fraction of that formed by the perfect duplex.

For what follows, it is assumed that conditions can be found which allow oligonucleotides which have complements in the probe to be distinguisned from those which do not.

4.3 ARRAY FORMAT, CONSTRUCTION AND SIZE

To form an idea of the scale of the arrays needed

to analyse sequences of different complexity it is convenient to think of the array as a square matrix. All sequences of a given length can be represented just once in a matrix constructed by drawing four rows representing the four bases, followed by four similar columns. This produces a 4 x 4 matrix in which each of the 16 squares represents one of the 16 doublets. Four similar matrices, but one quarter the size, are then drawn within each of the original squares. This produces a 16 x 16 matrix containing all 256 tetranucleotide sequences. Repeating this process produces a matrix of any chosen depth, s, with a number of cells equal to 4^s. As discussed above, the choice of s is of great importance, as it determines the complexity of the sequence representation. As discussed below, s also determines the size of the matrix constructed, which must be very big for complex genomes. Finally, the length of the oligonucleotides determines the hybridisation conditions and their discriminating power as hybridisation probes.

s	4^s	Genomes $4^s \times 10$	Side of Matrix (pixel=100 μm)	Number of Sheets of film
8	65536			
9	262144			
10	1.0×10^6	cosmid	100 m	1
11	4.2×10^6			
12	1.7×10^7			
13	6.7×10^7	E.coli		
14	2.6×10^8	yeast	1.6 m	9
15	1.1×10^9			
16	4.2×10^9			
17	1.7×10^{10}			
18	6.7×10^{10}	human	25 m	2,500
19	2.7×10^{11}			
20	1.1×10^{12}		100 m	

The table shows the expected scale of the arrays needed to perform the first analysis of a few genomes. The examples were chosen because they are genomes which have either been sequenced by conventional procedures - the cosmid scale -, are in the process of being sequenced - the E. coli scale -, or for which there has been considerable discussion of the magnitude of the problem - the human scale. the table shows that the expected scale of the matrix approach is only a small fraction of the conventional approach. This is readily seen in the area of X-ray film that would be consumed. It is also evident that the time taken for the analysis would be only a small fraction of that needed for gel methods. The "Genomes" column shows the length of random sequence which would fill about 5% of cells in the matrix. This has been determined to be the optimum condition for the first step in the sequencing strategy discussed below. At this size, a high proportion of the positive signals would represent single occurrences of each oligomer, the conditions needed to compare two genomes for sequence differences.

5. REFINEMENT OF AN INCOMPLETE SEQUENCE

Reconstruction of a complex sequence produces a result in which the reconstructed sequence is interrupted at any point where an oligomer that is repeated in the sequence occurs. Some repeats are present as components of long repeating structures which form part of the structural organisation of the DNA, dispersed and tandum repeats in human DNA for example. But when the length of oligonucleotide used in the matrix is smaller than that needed to give totally unique sequence representation, repeats occur by chance. Such repeats are likely to be isolated. That is, the sequences surrounding the repeated oligomers are unrelated to each other. The gaps caused

by these repeats can be removed by extending the sequence to longer oligomers. In principle, those sequences shown to be repeated by the first analysis, using an array representation of all possible oligomers, could be resynthesised with an extension at each end. For each repeated oligomer, there would be 4 x 4 = 16 oligomers in the new matrix. The hybridisation analysis would now be repeated until the sequence was complete. In practice, because the results of a positive signal in the hybridisation may be ambiguous, it may be better to adopt a refinement of the first result by extending all sequences which did not give a clear negative result in the first analysis. An advantage of this approach is that extending the sequence brings mismatches which are close to the ends in the shorter oligomer, closer to the centre in the extended oligomer, increasing the discriminatory power of duplex formation.

5.1 A HYPOTHETICAL ANALYSIS OF THE SEQUENCE OF BACTERIOPHAGE λ DNA

Lambda phage DNA is 48,502 base pairs long. Its sequence has been completely determined, we have treated one strand of this as a test case in a computer simulation of the analysis. The table shows that the appropriate size of oligomer to use for a sequence of this complexity is the 10-mer. With a matrix of 10-mers, the size was 1024 lines square. After "hybridisation" of the lambda 10-mers in the computer, 46,377 cells were positive, 1957 had double occurrences, 75 triple occurrences, and three quadruple occurrences. These 46,377 positive cells represented known sequences, determined from their position in the matrix. Each was extended by four x one base at the 3' end and four x one base at the 5', end to give 16 x 46,377 = 742,032 cells. This extended set reduced the

number of double occurrences to 161, a further 16-fold extension brought the number down to 10, and one more provided a completely overlapped result. Of course, the same end result of a fully overlapped sequence could be achieved starting with a 4^{16} matrix, but the matrix would be 4000 times bigger than the matrix needed to represent all 10-mers, and most of the sequence represented on it would be redundant.

5.2 LAYING DOWN THE MATRIX

The method described here envisages that the matrix will be produced by synthesising oligonucleotides in the cells of an array by laying down the precursors for the four bases in a predetermined pattern, an example of which is described above. Automatic equipment for applying the precursors has yet to be developed, but there are obvious possibilites; it should not be difficult to adapt a pen plotter or other computer-controlled printing device to the purpose. The smaller the pixel size of the array the better, as complex genomes need very large numbers of cells. However, there are limits to how small these can be made. 100 microns would be a fairly comfortable upper limit, but could probably not be achieved on paper for reasons of texture and diffusion. On a smooth impermeable surface, such as glass, it may be possible to achieve a resolution of around 10 microns, for example by using a laser typesetter to preform a solvent repellant grid, and building the oligonucleotides in the exposed regions. One attractive possibility, which allows adaptation of present techniques of oligonucleotide synthesis, is to sinter microporous glass in microscopic patches onto the surface of a glass plate. Laying down very large number of lines or dots could take a long time, if the printing mechanism were slow. However, a low cost ink-

jet printer can print at speeds of about 10,000 spots per second. With this sort of speed, 10^8 spots could be printed in about three hours.

5.3 OLIGONUCLEOTIDE SYNTHESIS

There are several methods of synthesising oligonucleotides. Most methods in current use attach the nucleotides to a solid support of controlled pore size glass (CPG) and are suitable for adaptation to synthesis on a glass surface. Although we know of no description of the direct use of oligonucleotides as hybridisation probes while still attached to the matrix on which they were synthesised, there are reports of the use of oligonucleotides as hybridisation probes on solid supports to which they were attached after synthesis. PCT Application WO 85/01051 describes a method for synthesising oligonucleotides tethered to a CPG column. In an experiment performed by us, CPG was used as the support in an Applied Bio-sytems oligonucleotide synthesiser to synthesise a 13-mer complementary to the left hand cos site of phage lambda. The coupling steps were all close to theoretical yield. The first base was stably attached to the support medium through all the synthesis and deprotection steps by a covalent link.

5.4 ANALYSING SEVERAL SEQUENCES SIMULTANEOUSLY

The method of this invention can be used to analyse several polynucleotide sequences simultaneously. To achieve this, the oligonucleotides may be attached to the support in the form of (for example) horizontal stripes. A technique for doing this is described in Example 3 below. Each DNA sample to be analysed is labelled and applied to the surface carrying the oligonucleotides in the form of a stripe (e.g. vertical) orthogonal to the oligonucleotide

stripes of the array. Hybridisation is seen at the intersections between oligonucleotide stripes and stripes of test sequence where there is homology between them.

Where sequence variations are known, an advantage of using this technique is that many different mutations can be probed simultaneously by laying down stripes corresponding to each allelic variant. With a density of one oligonucleotide per mm, and one "individual" per 5 mm, it should be possible to analyse 2000 loci on a plate 100 mm square. Such a high density of information, where the oligonucleotides do identify specific alleles, is not available by other techniques.

6. PROBES, HYBRIDISATION AND DETECTION

The yield of oligonucleotides synthesised on microporous glass is about 30 µmol/g. A patch of this material 1 micron thick by 10 microns square would hold $\approx 3 \times 10^{-12}$ µmol, equivalent to about 2 g of human

DNA. The hybridisation reaction could therefore be carried out with a very large excess of the bound oligonucleotides over that in the probe. So it should be possible to design a system capable of distinguishing between hybridisation involving single and multiple occurrances of the probe sequence, as yield will be proportional to concentration at all stages in the reaction.

The polynucleotide sequence to be analysed may be of DNA or RNA. To prepare the probe, the polynucleotide may be degraded to form fragments. Preferably it is degraded by a method which is as random as possible, to an average length around the chosen length s of the oligonucleotides on the support, and oligomers of exact length s selected by electrophoresis on a sequencing gel. The probe is then labelled. For example, oligonucleotides of length s may be end labelled. If labelled with ^{32}P, the radioactive yield of any individual s-mer even from total human DNA could be more than 10^4 dpm/mg of total DNA. For detection, only a small fraction of this is needed in a patch 10-100 microns square. This allows hybridisation conditions to be chosen to be close to the Tm of duplexes, which decreases the yield of hybrid and decreases the rate of formation, but increases the discriminating power. Since the bound oligonucleotide is in excess, signal need not be a problem even working close to equilibrium.

Hybridisation conditions can be chosen to be those known to be suitable in standard procedures used to hybridise to filters, but establishing optimum conditions is important. In particular, temperature needs to be controlled closely, preferably to better than $\pm 0.5^{\circ}C$. Particularly when the chosen length of the oligonucleotide is small, the analysis needs to be

able to distinguish between slight differences of rate and/or extent of hybridisation. The equipment may need to be programmed for differences in base composition between different oligonucleotides. In constructing the array, it may be preferable to partition this into sub-matrices with similar base compositions. This may make it easier to define the Tm which may differ slightly according to the base composition.

The choice of hybridisation solvent is significant. When 1M NaCl is used, G:C base pairs are more stable than A:T base pairs. Double stranded oligonucleotides with a high G+Ccontent have a higher Tm than corresponding oligonucleotides with a high A+T content. This discrepancy can be compensated in various ways: the amount of oligonucleotide laid down on the surface of the support can be varied depending on its nucleotide composition; or the computer used to analyse the data can be programmed to compensate for variations in nucleotide composition. A preferred method, which can be used either instead of or in addition to those already mentioned, is to use a chaotropic hybridisation solvent, for example a quarternary or tertiary amine as mentioned above. Tetramethylammoniumchloride (TMACl) has proved particularly suitable, at concentrations in the range 2 M to 5.5 M. At TMACl concentrations around 3.5 M to 4 M, the T_m dependence on nucleotide composition is greatly reduced.

The nature of the hybridisation salt used also has a major effect on the overall hybridisation yield. Thus, the use of TMACl at concentrations up to 5 M can increase the overall hybridisation yield by a factor of 30 or more (the exact figure depending to some extent on nucleotide composition) in comparison with hybridisation using 1M NaCl. Manifestly, this has important implications; for example the amount of

probe material that needs to be used to achieve a given signal can be much lower.

Autoradiography, especially with ^{32}P causes image degradation which may be a limiting factor determining resolution; the limit for silver halide films is around 25 microns. Obviously some direct detection system would be better. Fluorescent probes are envisaged; given the high concentration of the target oligonucleotides, the low sensitivity of fluorescence may not be a problem.

We have considerable experience of scanning autoradiographic images with a digitising scanner. Our present design is capable of resolution down to 25 microns, which could readily be extended down to less than present application, depending on the quality of the hybridisation reaction, and how good it is at distinguishing absence of a sequence from the presence of one or more. Devices for measuring astronomical plates have an accuracy around 1 μ. Scan speeds are such that a matrix of several million cells can be scanned in a few minutes. Software for the analysis of the data is straight-forward, though the large data sets need a fast computer.

Experiments presented below demonstrate the feasibility of the claims.

Commercially available microscope slides (BDH Super Premium 76 x 26 x 1 mm) were used as supports. These were derivatised with a long aliphatic linker that can withstand the conditions used for the

deprotection of the aromatic heterocyclic bases, i.e. 30% NH_3 at 55° for 10 hours. The linker, bearing a hydroxyl group which serves as a starting point for the subsequent oligonucleotide, is synthesised in two steps. The slides are first treated with a 25% solution of 3-glycidoxypropyltriethoxysilane in xylene containing several drops of Hunig's base as a catalyst. The reaction is carried out in a staining jar, fitted with a drying tube, for 20 hours at 90°C. The slides are washed with MeOH, Et_2O and air dried. Then neat hexaethylene glycol and a trace amount of conc. sulphuric acid are added and the mixture kept at 80° for 20 hours. The slides are washed with MeOH, Et_2O, air dried and stored desiccated at -20° until use. This preparative technique is described in British Patent Application 8822228.6 filed 21 September 1988.

The oligonucleotide synthesis cycle is performed as follows:

The coupling solution is made up fresh for each step by mixing 6 vol. of 0.5M tetrazole in anhydrous acetonitrile with 5 vol. of a 0.2M solution of the required beta-cyanoethylphosphoramidite. Coupling time is three minutes. Oxidation with a 0.1M solution of I_2 in THF/pyridine/H_2O yields a stable phosphotriester bond. Detritylation of the 5' end with 3% trichloroacetic acid in dichloromethane allows further extension of the oligonucleotide chain. There was no capping step since the excess of phosphoramidites used over reactive sites on the slide was large enough to drive the coupling to completion. After the synthesis is completed, the oligonucleotide is deprotected in 30% NH_3 for 10 hours at 55°. The chemicals used in the coupling step are moisture-sensitive, and this critical step must be performed under anhydrous conditions in a sealed container, as follows. The shape of the patch

to be synthesised was cut out of a sheet of silicone rubber (76 x 26 x 0.5 mm) which was sandwiched between a microscope slide, derivatised as described above, and a piece of teflon of the same size and thickness. To this was fitted a short piece of plastic tubing that allowed us to inject and withdraw the coupling solution by syringe and to flush the cavity with Argon. The whole assembly was held together by fold-back paper clips. After coupling the set-up was disassembled and the slide put through the subsequent chemical reactions (oxidation with iodine, and detritylation by treatment with TCA) by dipping it into staining jars.

EXAMPLE 1.

As a first example we synthesised the sequences oligo-dT_{10}-oligo-dT_{14} on a slide by gradually decreasing the level of the coupling solution in steps 10 to 14. Thus the 10-mer was synthesised on the upper part of the slide, the 14-mer at the bottom and the 11, 12 and 13-mers were in between. We used 10 pmol oligo-dA_{12}, labelled at the 5' end with ^{32}P by the polynucleotide kinase reaction to a total activity of 1.5 million c.p.m., as a hybridisation probe. Hybridisation was carried out in a perspex (Plexiglas) container made to fit a microscope slide, filled with 1.2 ml of 1M NaCl in TE, 0.1% SDS, for 5 minutes at 20^{o}. After a short rinse in the same solution without oligonucleotide, we were able to detect more than 2000 c.p.s. with a radiation monitor. An autoradiograph showed that all the counts came from the area where the oligonucleotide had been synthesised, i.e. there was no non-specific binding to the glass or to the region that had been derivatised with the linker only. After partial elution in 0.1 M NaCl differential binding to the target is detectable, i.e. less binding to the shorter than the longer oligo-dT. By gradually heating the

slide in the wash solution we determined the T_{om} (mid-point of transition when 50% eluted) to be 33°. There were no counts detectable after incubation at 39°. The hybridisation and melting was repeated eight times with no diminution of the signal. The result is reproducible. We estimate that at least 5% of the input counts were taken up by the slide at each cycle.

EXAMPLE 2.

In order to determine whether we would be able to distinguish between matched and mismatched oligonucleotides we synthesised two sequences 3' CCC GCC GCT GGA (cosL) and 3' CCC GCC TCT GGA, which differ by one base at position 7. All bases except the seventh were added in a rectangular patch. At the seventh base, half of the rectangle was exposed in turn to add the two different bases, in two stripes. Hybridisation of cosR probe oligonucleotide (5' GGG CGG CGA CCT) (kinase labelled with ^{32}P to 1.1 million c.p.m., 0.1 M NaCl, TE, 0.1% SDS) was for 5 hours at 32°. The front of the slide showed 100 c.p.s. after rinsing. Autoradiography showed that annealing occurred only to the part of the slide with the fully complementary oligonucleotide. No signal was detectable on the patch with the mismatched sequence.

EXAMPLE 3.

For a further study of the effects of mismatches or shorter sequences on hybridisation behaviour, we constructed two arrays; one (a) of 24 oligonucleotides and the other (b) of 72 oligonucleotides.

These arrays were set out as shown in Table 1(a) and 1(b). The masks used to lay down these arrays were different from those used in previous experiments. Lengths of silicone rubber tubing (1mm o.d.) were glued with silicone rubber cement to the surface of plain microscope slides, in the form of a "U". Clamping

these masks against a derivatised microscope slide produced a cavity into which the coupling solution was introduced through a syringe. In this way only the part of the slide within the cavity came into contact with the phosphoramidite solution. Except in the positons of the mismatched bases, the arrays listed in Table 1 were laid down using a mask which covered most of the width of the slide. Off-setting this mask by 3mm up or down the derivatised slide in subsequent coupling reactions produced the olignucleotides truncated at the 3' or 5' ends.

For the introduction of mismatches a mask was used which covered half (for array (a)) or one third (for array (b)) of the width of the first mask. The bases at positions six and seven were laid down in two or three longitudinal stripes. This led to the synthesis of oligonucleotides differing by one base on each half (array (a)) or third (array (b)) of the slide. In other positions, the sequences differed from the longest sequence by the absence of bases at the ends.

In array (b), there were two columns of sequences between those shown in Table 1(b), in which the sixth and seventh bases were missing in all positions, because the slide was masked in a stripe by the silicone rubber seal. Thus there were a total of 72 different sequences represented on the slide in 90 different positions.

The 19-mer 5' CTC CTG AGG AGA AGT CTG C was used for hybridisation (2 million cpm, 1.2 ml 0.1M NaCl in TE, 0.1% SDS, 20°C).

The washing and elution steps were followed by autoradiography. The slide was kept in the washing solution for 5 min at each elution step and then exposed (45 min, intensified). Elution temperatures were 23, 36, 42, 47, 55 and 60°C respectively.

As indicated in the table, the oligonucleotides showed different melting behaviour. Short oligonucleotides melted before longer ones, and at 55°C, only the perfectly matched 19-mer was stable, all other oligonucleotides had been eluted. Thus the method can differentiate between a 18-mer and a 19-mer which differ only by the absence of one base at the end. Mismatches at the end of the oligonucleotides and at internal sites can all be melted under conditions where the perfect duplex remains.

Thus we are able to use very stringent hybridisation conditions that eliminate annealing to mismatch sequences or to oligonucleotides differing in length by as little as one base. No other method using hybridisation of oligonucleotides bound to the solid supports is so sensitive to the effects of mismatching.

EXAMPLE 4.

To test the application of the invention to diagnosis of inherited diseases, we hybridised the array (a), which carries the oligonucleotide sequences specific for the wild type and the sickle cell mutations of the β -globin gene, with a 110 base pair fragment of DNA amplified from the β -globin gene by means of the polymerase chain reaction (PCR). Total DNA from the blood of a normal individual (1 microgram) was amplified by PCR in the presence of appropriate primer oligonucleotides. The resulting 110 base pair fragment was purified by electrophoresis through an agarose gel. After elution, a small sample (ca. 10^{-3} picogram) was labelled by using α - ^{32}P-dCTP (50 microCurie) in a second PCR reaction. This PCR contained only the upstream priming oligonucleotide. After 60 cycles of amplification with an extension time of 9 min. the product was removed from precursors by gel filtration. Gel electrophoresis of the radioactive product showed a

major band corresponding in length to the 110 base fragment. One quarter of this product (100,000 c.p.m. in 0.9 M NaCl, TE, 0.1% SDS) was hybridised to the array (a). After 2 hours at 30°C ca. 15000 c.p.m. had been taken up. The melting behaviour of the hybrids was followed as described for the 19-mer in example 3, and it was found that the melting behaviour was similar to that of the oligonucleotide. That is to say, the mismatches considerably reduced the melting temperature of the hybrids, and conditions were readily found such that the perfectly matched duplex remained whereas the mismatched duplexes had fully melted.

Thus the invention can be used to analyse long fragments of DNA as well oligonucleotides, and this example shows how it may be used to test nucleic acid sequences for mutations. In particular it shows how it may be applied to the diagnosis of genetic diseases.

EXAMPLE 5.

To test an automated system for laying down the precursors, the cosL oligonucleotide was synthesised with 11 of the 12 bases added in the way described above. For the addition of the seventh base, however, the slide was transferred into an Argon filled chamber containing a pen plotter. The pen of the plotter had been replaced by a component, fabricated from Nylon, which had the same shape and dimensions as the pen, but which carried a polytetrafluoroethylene (PTFE) tube, through which chemicals could be delivered to the surface of the glass slide which lay on the bed of the plotter. A microcomputer was used to control the plotter and the syringe pump which delivered the chemicals. The pen, carrying the delivery tube from the syringe, was moved into position above the slide, the pen was lowered and the pump activated to lay down coupling solution. Filling the pen successively with

G, T and A phosphoramidite solutions an array of twelve spots was laid down in three groups of four, with three different oligonucleotide sequences. After hybridisation to cosR, as described in Example 2, and autoradiography, signal was seen only over the four spots of perfectly matched oligonucleotides, where the dG had been added.

EXAMPLE 6

This example demonstrates the technique of analysing several DNA sequences simultaneously. Using the technique described in Example 3, a slide was prepared bearing six parallel rows of oligonucleotides running along its length. These comprised duplicate hexadecamer sequences corresponding to antisense sequences of the β-globin wild-type (A), sickle cell (S) and C mutations.

Clinical samples of AC, AS and SS DNA were procured. Three different single-stranded probes of 110 nt length with approx. 70,000 c.p.m. in 100 µl 1M NaCl, TE pH 7.5, 0.1% SDS, viz AC, AS, and SS DNA were prepared. Radiolabelled nucleotide was included in the standard PCR step yielding a double-stranded labelled fragment. It was made single-stranded with Bacteriophage λ exonuclease that allowed to selectively digest one strand bearing a 5′ phosphate. This was made possible by phosphorylating the downstream primer with T4 Polynucleotide kinase and ('cold') ATP prior to PCR. These three probes were applied as three stripes orthogonal to the surface carrying the six oligonucleotide stripes. Incubation was at 30°C for 2 hours in a moist chamber. The slide was then rinsed at ambient temperature, then 45°C for 5 minutes and exposed for 4 days with intensification. The genotype of each clinical sample was readily determined from the autoradiographic signals at the points of intersection.

EXAMPLE 7

A plate was prepared whose surface carried an array of all 256 octapurines. That is to say, the array comprised 256 oligonucleotides each consisting of a different sequence of A and G nucleotides. This array was probed with a mixture comprising all 256 octapyrimidines, each end labelled by means of polynucleotide kinase and γ-^{32}P-ATP. Hybridisation was performed for 6 - 8 hours at 4°C.

In consecutive experiments the hybridisation solvent was changed through the series 1M NaCl (containing 10mM Tris.HCl pH 7.5, 1mM EDTA, 7% sarcosine) and 2M, 2.5M, 3M, 3.5M, 4M, 4.5M, 5M and 5.5M TMACl (all containing 50mM Tris.HCl pH 8.0, 2mM EDTA, SDS at less than 0.04 mg/ml). The plate was rinsed for 10 minutes at 4°C in the respective solvent to remove only loosely matched molecules, sealed in a plastic bag and exposed to a PhorphorImager storage phosphor screen at 4°C overnight in the dark.

The following table quotes relative signal intensities, at a given salt concentration, of hybrids formed with oligonucleotides of varying a content. In this table, the first row refers to the oligonucleotide GGGGGGGG, and the last row to the oligonucleotide AAAAAAAA. It can be seen that the difference in response of these two oligonucleotides is marked in 1M NaCl, but much less marked in 3M or 4M TMACl.

Relative Intensities at given Salt Concentration

Solvent	Number of A's		
	0	4	8
1M NaCl	100	30	20
2M TMACl	100	70	30
3M TMACl	70	100	40
4M TMACl	60	100	40

The following table indicates relative signal intensities obtained, with octamers containing 4A's and 4G's, at different hybridisation salt concentrations. It can be seen that the signal intensity is dramatically increased at higher concentrations of TMACl.

Relative Intensities at different Salt Concentrations

Solvent	Yield of hybrid
1M NaCl	100
2M TMACl	200
3M TMACl	700
4M TMACl	2000

In conclusion, we have demonstrated the following:

1. It is possible to synthesise oligonucleotides in good yield on a flat glass plate.

2. Multiple sequences can be synthesised on the sample in small spots, at high density, by a simple manual procedure, or automatically using a computer controlled device.

3. Hybridisation to the oligonucleotides on the plate can be carried out by a very simple procedure. Hybridisation is efficient, and hybrids can be detected by a short autoradiographic exposure.

4. Hybridisation is specific. There is no detectable signal on areas of the plate where there are no oligonucleotides. We have tested the effects of mismatched bases, and found that a single mismatched base at any position in oligonucleotides ranging in length from 12-mer to 19-mer reduces the stability of the hybrid sufficiently that the signal can be reduced to a very low level, while retaining significant hybridisation to the perfectly matched hybrid.

5. The oligonucleotides are stably bound to the glass and plates can be used for hybridisation repeatedly.

The invention thus provides a novel way of analysing nucleotide sequences, which should find a wide range of application. We list a number of potential applications below:

Small arrays of oligonucleotides as fingerprinting and mapping tools

Analysis of known mutations including genetic diseases.

Example 4 above shows how the invention may be used to analyse mutations. There are many applications for such a method, including the detection of inherited diseases.

Genomic fingerprinting.

In the same way as mutations which lead to disease can be detected, the method could be used to detect point mutations in any stretch of DNA. Sequences are now available for a number of regions containing the base differences which lead to restriction fragment length polymorphisms (RFLPs). An array of oligonucleotides representing such polymorphisms could be made from pairs of oligonucleotides representing the two allelic restriction sites. Amplification of the sequence containing the RFLP, followed by hybridisation to the plate, would show which alleles were present in the test genome. The number of oligonucleotides that could be analysed in a single analysis could be quite large. Fifty pairs made from selected alleles would be enough to give a fingerprint unique to an individual.

Linkage analysis.

Applying the method described in the last paragraph to a pedigree would pinpoint recombinations. Each pair of spots in the array would give the information that is seen in the track of the RFLP analysis, using gel electrophoresis and blotting, that is now routinely used for linkage studies. It should be possible to analyse many alleles in a single analysis, by hybridisation to an array of allelic pairs of oligonucleotides, greatly simplifying the methods used to find linkage between a DNA polymorphism and phenotypic marker such as a disease gene.

The examples above could be carried out using the

method we have developed and confirmed by experiments.

Large arrays of oligonucleotides as sequence reading tools.

We have shown that oligonucleotides can be synthesised in small patches in precisely determined positions by one of two methods: by delivering the precursors through the pen of a pen-plotter, or by masking areas with silicone rubber. It is obvious how a pen plotter could be adapted to synthesise large arrays with a different sequence in each position. For some applications the array should be a predetermined, limited set; for other applications, the array should comprise every sequence of a predetermined length. The masking method can be used for the latter by laying down the precursors in a mask which produces intersecting lines. There are many ways in which this can be done and we give one example for illustration:

1. The first four bases, A, C, G, T, are laid in four broad stripes on a square plate.
2. The second set is laid down in four stripes equal in width to the first, and orthogonal to them. The array is now composed of all sixteen dinucleotides.
3. The third and fourth layers are laid down in four sets of four stripes one quarter the width of the first stripes. Each set of four narrow stripes runs within one of the broader stripes. The array is now composed of all 256 tetranuclectides.
4. The process is repeated, each time laying down two layers with stripes which are one quarter the width of the previous two layers. Each layer added increases the length of the oligonucleotides by one base, and the number of different oligonucleotide sequences by a factor of four.

The dimensions of such arrays are determined by the width of the stripes. The narrowest stripe we

have laid is 1mm, but this is clearly not the lowest limit.

There are useful applications for arrays in which part of the sequence is predetermined and part made up of all possible sequences. For example:

Characterising mRNA populations.

Most mRNAs in higher eukaryotes have the sequence AAUAAA close to the 3' end. The array used to analyse mRNAs would have this sequence all over the plate. To analyse a mRNA population it would be hybridised to an array composed of all sequences of the type N_m AATAAA N_n. For m + n = 8, which should be enough to give a unique oligonucleotide address to most of the several thousand mRNAs that is estimated to be present in a source such as a mammalian cell, the array would be 256 elements square. The 256 x 256 elements would be laid on the AATAAA using the masking method described above. With stripes of around 1mm, the array would be ca. 256mm square.

This analysis would measure the complexity of the mRNA population and could be used as a basis for comparing populations from different cell types. The advantage of this approach is that the differences in the hybridisation pattern would provide the sequence of oligonucleotides that could be used as probes to isolate all the mRNAs that differed in the populations.

Sequence determination.

To extend the idea to determine unkown sequences, using an array composed of all possible oligonucleotides of a chosen length, requires larger arrays than we have synthesised to date. However, it is possible to scale down the size of spot and scale up the numbers to those required by extending the methods we have developed and tested on small arrays. Our experience shows that the method is much simpler in operation than the gel based methods.

TABLE 1

For Examples 3 and 4 array (a) was set out as follows:

20	GAG GAC TCC TCT ACG		20	GAG GAC aCC TCT ACG	
36	GAG GAC TCC TCT GAC G		20	GAC GAC aCC TCT GAC G	
36	GAG GAC TCC TCT AGA CG		20	GAC GAC aCC TCT AGA CG	
47	GAG GAC TCC TCT CAG ACG		36	GAG GAC aCC TCT CAG ACG	
60	GAG GAC TCC TCT TCA GAC G		47	GAG GAC aCC TCT TCA GAC G	
56	.AG GAC TCC TCT TCA GAC G		42	.AG GAC aCC TCT TCA GAC G	
56	..G GAC TCC TCT TCA GAC G		42	..G GAC aCC TCT TCA GAC G	
47	... GAC TCC TCT TCA GAC G		42	... GAC aCC TCT TCA GAC G	
42	... .AC TCC TCT TCA GAC G		36	... .AC aCC TCT TCA GAC G	
36	... ..C TCC TCT TCA GAC G		36	... ..C aCC TCT TCA GAC G	
36	... ... TCC TCT TCA GAC G		36	... ... aCC TCT TCA GAC G	
36	... ... .CC TCT TCA GAC G		36	... ... .CC TCT TCA GAC G	

For example 3 array (b) was set out as follows:

20 GAG GAt TC	20 GAG GAC TC	20 GAG GAC aC	
20 GAG GAt TCC	20 GAG GAC TCC	20 GAG GAC aCC	
20 GAG GAt TCC T	20 GAG GAC TCC T	20 GAG GAC aCC T	
20 GAG GAt TCC TC	20 GAG GAC TCC TC	20 GAG GAC aCC TC	
20 GAG GAt TCC TCT	20 GAG GAC TCC TCT	20 GAG GAC aCC TCT	
20 GAG GAt TCC TCT T	20 GAG GAC TCC TCT T	20 GAG GAC aCC TCT T	
20 GAG GAt TCC TCT TC	20 GAG GAC TCC TCT TC	20 GAG GAC aCC TCT TC	
20 GAG GAt TCC TCT TCA	20 GAG GAC TCC TCT TCA	20 GAG GAC aCC TCT TCA	
32 GAG GAt TCC TCT TCA G	42 GAG GAC TCC TCT TCA G	20 GAG GAC aCC TCT TCA G	
32 GAG GAt TCC TCT TCA GA	47 GAG GAC TCC TCT TCA GA	32 GAG GAC aCC TCT TCA GA	
42 GAG GAt TCC TCT TCA GAC	52 GAG GAC TCC TCT TCA GAC	42 GAG GAC aCC TCT TCA GAC	
52 GAG GAt TCC TCT TCA GAC G	60 GAG GAC TCC TCT TCA GAC G	52 GAG GAC aCC TCT TCA GAC G	
42 .AG GAt TCC TCT TCA GAC G	52 .AG GAC TCC TCT TCA GAC G	42 .AG GAC aCC TCT TCA GAC G	
42 ..G GAt TCC TCT TCA GAC G	52 ..G GAC TCC TCT TCA GAC G	42 ..G GAC aCC TCT TCA GAC G	
37 ... GAt TCC TCT TCA GAC G	47 ... GAC TCC TCT TCA GAC G	37 ... GAC aCC TCT TCA GAC G	
32 ... .At TCC TCT TCA GAC G	42 ... .AC TCC TCT TCA GAC G	32 ... .AC aCC TCT TCA GAC G	
32 ... ..t TCC TCT TCA GAC G	42 ... ..C TCC TCT TCA GAC G	32 ... ..C aCC TCT TCA GAC G	
32 ... ... TCC TCT TCA GAC G	32 ... ... TCC TCT TCA GAC G	32 ... ... aCC TCT TCA GAC G	

Between the three columns of array (b) listed above, were two columns, in which bases 6 and 7 from the left were missing in every line. These sequences all melted at 20 or 32 degrees. (a,t) mismatch base (.) missing base.